Appl. Statist. (2005)
54, *Part* 1, *pp.* 1–11

A survival-adjusted quantal response test for comparing tumour incidence rates

Shyamal D. Peddada, Gregg E. Dinse and Joseph K. Haseman

National Institute of Environmental Health Sciences, Research Triangle Park, USA

[Received July 2003. Final revision November 2003]

Summary. The paper presents a case-study of skin fibromas among male rats in the 2-year cancer bioassay of methyleugenol that was conducted by the US National Toxicology Program (NTP). In animal carcinogenicity experiments such as this one, tumour rates are often compared with the Cochran–Armitage (CA) trend test. The operating characteristics of the CA test, however, can be adversely affected by survival differences across groups and by the assumed dose metric. Survival-adjusted generalizations of the CA test have been proposed, but they are still sensitive to the choice of scores that are assigned to the dose groups. We present an alternative test, which outperforms the survival-adjusted CA test which is currently used by the NTP to compare incidence rates. Simulated data from a wide range of realistic situations show that the operating characteristics of the test proposed are superior to those of the NTP's survival-adjusted CA test, especially for rare tumours and wide logarithmic spacings of the dose metric.

Keywords: Animal carcinogenicity study; Cancer bioassay; Cochran–Armitage trend test; National Toxicology Program; Order-restricted inference; Poly-3 test; Tumour onset

1. Introduction

The National Toxicology Program (NTP) is the US Government's premier toxicological research organization. As part of its mission, the NTP conducts 2-year rodent bioassays to investigate possible carcinogenic effects of various chemicals. We recently had occasion to examine the data on skin fibromas among male rats in the NTP's methyleugenol bioassay (National Toxicology Program, 2000). Our paper presents a case-study of this data set, and we provide methods that others might find useful when analysing similar data sets.

The NTP methyleugenol study involved 50 male rats in each of four treatment groups: a control group and three dose groups (Table 1). Researchers were interested in determining whether certain tumour rates increased with the dose of methyleugenol. The numbers of rats with skin fibromas in the four groups were 1, 9, 8 and 5. These observed incidences varied with dose, but not monotonically. The corresponding 2-year survival rates were 40%, 33%, 32% and 0%. In case the downturn in observed incidence was attributable, at least partly, to the dose-related decrease in survival, the NTP calculated survival-adjusted skin fibroma rates of 2.4%, 22.3%, 20.6% and 15.3% respectively. Thus, adjusting for survival lessened, but did not eliminate, the apparent downturn in observed tumour rates. Probably as a result of this downturn, the NTP's survival-adjusted linear trend test failed to detect what many researchers might view as a trend in skin fibroma rates.

Address for correspondence: Shyamal Peddada, Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA.
E-mail: peddada@niehs.nih.gov

 0035–9254/05/54000

Table 1. Skin fibroma and survival data for 200 male rats in the NTP bioassay of methyleugenol†

Dose (mg kg^{-1})	*Death times (days) and tumour response information*
0	344, 521, 529, 553, 564, 588, 603, 610 (2), 614, 617, 626, 639, 642, 652, 661, 662, 668, 704 (2), 707 (2), 712, 714, 715, 718, 721, 722 (2), 729, 730 (19), 730‡
37	403, 406, 431, 535‡, 542‡, 572, 575, 579 (2), 589, 591, 596, 602, 606, 625, 638, 639, 650, 660, 661‡, 664, 673‡, 674, 680, 695, 703, 704 (3), 712, 718, 718‡ (2), 725, 730 (13), 730‡ (3)
75	15, 438, 467, 502, 521, 522, 568, 572, 582, 595, 596, 601, 610, 619‡, 630, 638, 639‡, 642 (2), 651, 658, 659‡, 661, 664, 673‡, 674, 692, 695 (2), 695‡, 696 (2), 706, 709, 712‡, 730 (13), 730‡ (2)
150	337, 409, 457, 467, 495, 502, 523, 546, 547, 568, 575, 583, 584, 598 (2), 600, 602, 607‡, 610, 614, 621, 625, 633‡, 638, 642 (3), 646, 648, 650, 654, 658, 659, 660 (2), 660‡ (2), 661, 669, 670, 680 (2), 683‡, 684, 688 (2), 699, 700, 704, 712

†The terminal sacrifice was performed at 730 days, and the number of tied death times is given in parentheses.
‡Presence of a skin fibroma at death.

Historically, the NTP has employed a form of the linear trend test of Cochran (1954) and Armitage (1955) to assess dose-related increases in the proportion of animals that are expected to develop a tumour. Two important issues in the application of the standard Cochran–Armitage (CA) trend test are differences in survival between the treatment groups and the spacing of the dose metric. The CA test assumes that all animals in the same treatment group are at equal risk of developing a tumour. This assumption is usually violated, however, because the risk of developing a tumour increases with age and generally animals die at various times during the study. Differential mortality within a treatment group can decrease the efficiency of the CA test and survival patterns that vary across treatment groups, as was the case in the methyleugenol bioassay, can produce biased inferences.

Doses are often logarithmically spaced, typically increasing by a multiple of 2–5; thus, the CA trend test can give the observed tumour response at the top dose considerable weight in the statistical analysis. The high dose occasionally shows a downturn in tumour response, however, as occurred in the methyleugenol study. A downturn, which may reflect chemical toxicity and/or saturation of the metabolic pathways by which the effective (internal) dose is delivered, is inconsistent with a linear model. In such cases, the CA linear trend test has reduced power. Thus, it is desirable to develop a test that adjusts for differences in survival between groups and is less affected by downturns in response at the high dose.

The NTP currently uses a survival-adjusted version of the CA test known as the 'poly-3' test (Bailer and Portier, 1988). We introduce a new test for comparing incidence rates, which retains the simplicity of the poly-3 test but improves on its performance. When applied to the methyleugenol data, our test reveals a statistically significant trend in the incidence of skin fibromas. The test proposed extends the order-restricted trend test of Peddada *et al.* (2001) to accommodate differential mortality through a poly-3 survival adjustment. Our test is easy to implement and yet it generally outperforms the usual poly-3 test. When evaluating competing methods, we do not consider those which require extra data, such as interim sacrifices or cause-of-death determinations, so that our results will apply in the broadest range of real situations. In fact, we focus entirely on our test and the poly-3 adjusted CA test, as the latter is arguably the most widely used of the competing procedures.

Section 2 gives some background material, defines our notation and outlines previous approaches. Section 3 describes the test proposed and Section 4 provides further discussion

of the methyleugenol analysis. Section 5 compares our test and the poly-3 adjusted CA test, based on simulated data from a wide range of realistic situations.

2. Background

A variety of survival-adjusted methods have been proposed to solve the problems that are caused by differential mortality. Depending on the end point of interest, these methods typically require either non-standard data or unverifiable assumptions. Some common end points include the rate of death with a tumour, the rate of death caused by a tumour, the tumour prevalence rate and the tumour incidence rate. Typical types of non-standard information and simplifying assumptions include multiple sacrifices, cause-of-death assignments, tumour lethality restrictions, parametric distributions and functional constraints. For an overview of many of these survival-adjusted methods and related issues, see Dinse (1998).

Regarding the choice of the most appropriate end point, McKnight and Crowley (1984) argued convincingly in favour of the tumour incidence rate, which corresponds to the hazard function for time to the onset of a tumour. As a function of time, the incidence rate automatically accounts for differential mortality and, as the rate at which new tumours occur, it is a logical measure of carcinogenesis. The primary drawback is that most tumours are not observable in live animals. Thus, tumour incidence rates typically cannot be assessed directly and their analysis usually depends on collecting special data or making simplifying assumptions.

All carcinogenicity studies should provide observations on each animal's time of death and tumour status at death. In contrast, however, it may be cost prohibitive to conduct studies with multiple sacrifice times, and it may not always be possible to obtain data on the cause of death, as required, for example, by the approach of Peto *et al.* (1980). Furthermore, even when available, cause-of-death assignments are often unreliable (see, for example, Kodell *et al.* (1995)). In the absence of these types of additional information, tumour incidence analyses typically must make simplifying assumptions. In general, though, researchers would prefer to avoid assuming fixed levels of tumour lethality, particular parametric distributions or specific relationships between the functions of interest.

Motivated by these concerns, Bailer and Portier (1988) proposed a procedure which applies the CA test after reducing the sample sizes to adjust for survival (i.e. the time at risk). Their approach assumes that tumours occur at a rate that is proportional to a given power of time. Although the optimal power might vary with the type of tumour and other factors, Bailer and Portier examined a large set of historical data and suggested that a power of 3 might generally work well in practice. The resulting test, the so-called poly-3 test, has generated much interest. This poly-3 adjusted CA test, which we label CA_{P3}, permits valid inferences about the tumour incidence rates when its power assumption is satisfied but otherwise can produce misleading results (see, for example, Mancuso *et al.* (2002)). Other inferential procedures are available which do not require this power of 3 restriction (see, for example, Dinse (1991) and Dunson and Dinse (2001)), but they involve alternative assumptions or constraints. The main advantage of the poly-3 approach to survival adjustment is its simplicity.

2.1. Functions of interest

Let T denote the time to the first of two events, either tumour onset or death, and let $Y(t)$ be an indicator that is 1 if the (irreversible) tumour is present at time t and 0 otherwise. If T is a continuous random variable, the tumour incidence rate and the tumour-free death-rate can be expressed as the event-specific hazard functions

$$\lambda(t) = \lim_{\varepsilon \to 0} \{\Pr(t \leqslant T < t+\varepsilon, Y(T) = 1 | T \geqslant t)/\varepsilon\}$$

and

$$\beta(t) = \lim_{\varepsilon \to 0} \{\Pr(t \leqslant T < t+\varepsilon, Y(T) = 0 | T \geqslant t)/\varepsilon\}$$

respectively. Let π denote the expected proportion of animals that develop a tumour during the study, which can be expressed as the following function of the above two hazard rates:

$$\pi = \int_0^{t_s} \lambda(s) \exp\left[-\int_0^s \{\lambda(r) + \beta(r)\}\, dr \right] ds, \tag{1}$$

where t_s denotes the time at which the study ends (i.e. the terminal sacrifice time).

Let π_i, $\lambda_i(t)$ and $\beta_i(t)$ denote the expected tumour proportion, tumour incidence rate and tumour-free death-rate in the ith of k treatment groups ($i = 1, 2, \ldots, k$). We are interested in testing for a trend in tumour incidence rates, which translates into the null and ordered alternative hypotheses

$$\begin{aligned} H_0 &: \lambda_1(t) = \lambda_2(t) = \ldots = \lambda_k(t), \\ H_a &: \lambda_1(t) \leqslant \lambda_2(t) \leqslant \ldots \leqslant \lambda_k(t) \end{aligned} \tag{2}$$

respectively, where H_a involves at least one strict inequality. Expression (1) demonstrates that tests that are oriented towards the $\{\pi_i\}$ are not necessarily valid for comparing the $\{\lambda_i(t)\}$, as such tests can be confounded by differences between the $\{\beta_i(t)\}$. Our goal is to evaluate the operating characteristics of two quantal response trend tests with respect to the hypotheses in expression (2), which involve the $\{\lambda_i(t)\}$, as opposed to the usual hypotheses regarding the $\{\pi_i\}$.

2.2. Observed data

Suppose that there are n_i animals in the ith group ($i = 1, 2, \ldots, k$), each exposed to dose d_i of the treatment, where $d_1 < d_2 < \ldots < d_k$. Animals in the first group are unexposed controls ($d_1 = 0$). Let y_{ij} be an indicator that is 1 if the jth animal in the ith group has a tumour at death and is 0 otherwise. In addition, let $y_{i+} = \Sigma_{j=1}^{n_i} y_{ij}$ denote the number of animals in the ith group that die with a tumour, let $y_{++} = \Sigma_{i=1}^{k} y_{i+}$ denote the total number of animals dying with a tumour and let $n_+ = \Sigma_{i=1}^{k} n_i$ denote the total number of animals in the study.

2.3. Previous approaches

One of the original and most widely implemented trend tests was proposed by Cochran (1954) and Armitage (1955). Historically, in the context of carcinogenicity studies, the CA test has been used to compare groups with respect to the expected proportion of animals developing a tumour, i.e. the $\{\pi_i\}$ rather than the $\{\lambda_i(t)\}$. As mentioned earlier, however, the usual CA analysis can yield biased inferences when mortality patterns, i.e. the $\{\beta_i(t)\}$, differ across dose groups, as commonly occurs when the treatment is toxic.

For this reason, Bailer and Portier (1988) introduced the poly-3 survival adjustment. Briefly, they constructed a modified sample size n_i^* as the sum of animal-specific weights. Any animal that dies with a tumour, regardless of how early, or that survives until the end of the study, contributes a weight of $w = 1$. Otherwise, an animal that dies at time $t < t_s$ without a tumour contributes a fractional weight of $w = (t/t_s)^3$, as such an animal was not at risk of tumour onset for the entire study. Bailer and Portier then applied the CA test after substituting n_i^* for n_i ($i = 1, 2, \ldots, k$). Note that $\hat{\pi}_i = y_{i+}/n_i^*$ provides a survival-adjusted estimate of π_i. Bieler and

Williams (1993) corrected the variance to allow for the fact that n_i^* is not fixed. Bailer and Portier (1988) demonstrated that, although CA_{P3} is oriented towards comparing the $\{\pi_i\}$, it also provides an approximately valid comparison of the $\{\lambda_i(t)\}$ when the tumour incidence rates follow a Weibull model with a shape parameter of 3. However, simulation studies show that the CA_{P3} test is sensitive to certain departures from the underlying tumour incidence model (see, for example, Mancuso *et al.* (2002)).

Peddada *et al.* (2001) developed a general order-restricted methodology, which in the context of animal carcinogenicity studies can be applied to compare proportions of tumour-bearing animals. In the current situation, their basic test statistic reduces to

$$Z = \tilde{\pi}_k - \tilde{\pi}_1, \tag{3}$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_k$ are poly-3 adjusted isotonic estimates of the tumour proportions in the first and last dose groups respectively, i.e. the $\{\tilde{\pi}_i\}$ are the isotonic regression estimates (see, for example, Barlow *et al.* (1972)) obtained by applying the pool adjacent violators algorithm to the $\{y_{i+}\}$ with weights $\{n_i^*\}$. Peddada *et al.* (2001) used bootstrap techniques to obtain critical values for Z as follows. Combine the data from all k groups and use simple random sampling with replacement to assign n_i animals, each with its own poly-3-adjusted weight, to the ith group $(i = 1, \ldots, k)$. Compute the isotonic estimates of the $\{\pi_i\}$, assuming that they are monotone non-decreasing, and use equation (3) to calculate a Z^* for each bootstrap sample. Peddada *et al.* (2001) defined the α-level rejection region as $\{Z \geqslant z_{1-\alpha}\}$, where $z_{1-\alpha}$ is the $100(1-\alpha)$-percentile of the bootstrap distribution of the $\{Z^*\}$. In a related approach, Mancuso *et al.* (2001) also used isotonic regression and bootstrap techniques, but they focused on a modified CA_{P3} test statistic rather than Z. In this paper, we consider only one-sided tests of equality *versus* a non-decreasing trend. If required by a different application, it is straightforward to generalize the proposed methodology to handle other order restrictions along the lines of Peddada *et al.* (2001) with suitably calculated critical values. In the current application, however, extensions to more general hypotheses are not of interest.

The bootstrap procedures of Peddada *et al.* (2001) and Mancuso *et al.* (2001) assume that, under the null hypothesis, animals are exchangeable among the treatment groups. This exchangeability assumption can be violated in the presence of dose-related differential mortality, though, such as when the treatment is toxic (and reduces survival) but is not tumourigenic. Thus, since we are interested in situations where mortality varies with dose, as in the methyleugenol study, we do not include these bootstrap methods in our comparisons.

3. Method proposed

Bieler and Williams (1993) derived the following estimator for the variance of $\hat{\pi}_i = y_{i+}/n_i^*$ under the null hypothesis that $\pi_1 = \pi_2 = \ldots = \pi_k$:

$$\text{var}(\hat{\pi}_i) = S^2 n_i / n_i^{*2},$$

where $S^2 = \Sigma_{ij}(r_{ij} - \bar{r}_i)^2/(n_+ - k)$, $r_{ij} = y_{ij} - \hat{\pi} w_{ij}$, $\bar{r}_i = r_{i+}/n_i$, $\hat{\pi} = y_{++}/n_+^*$ and w_{ij} is the poly-3 weight for animal j in group i. Using this variance estimator, we propose the following statistic for testing the null hypothesis in expression (2) that $\lambda_1(t) = \lambda_2(t) = \ldots = \lambda_k(t)$:

$$W_{P3} = \frac{\tilde{\pi}_k - \tilde{\pi}_1}{S\sqrt{(n_1/n_1^{*2} + n_k/n_k^{*2})}}.$$

Our test statistic is similar to that of Peddada *et al.* (2001), but we base our rejection region on an asymptotic approximation rather than on a bootstrap distribution. The above test statistic

is also similar in spirit to one developed by Williams (1977) for testing simple ordering of normal means. In fact, we obtain an approximate null distribution for our test statistic by using ideas from Williams. Let $X_i \sim N(0,1)$ for $i = 1,\ldots,k$ and suppose that $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k$ denote the isotonized values of the $\{X_i\}$, using unit weights. Define

$$V = \frac{\hat{X}_k - \hat{X}_1}{\sqrt{2}}.$$

The distribution of V can be easily simulated and its percentiles can be used to approximate the corresponding percentiles for the distribution of W_{P3}. Thus, we reject H_0 if $W_{P3} \geqslant v_{1-\alpha}$, where $v_{1-\alpha}$ is the $100(1-\alpha)$-percentile of the simulated distribution of V.

4. Analysis of methyleugenol data

Methyleugenol is widely used as a fragrance and as a flavouring agent. Groups of 50 male rats received methyleugenol in 0.5% methylcellulose by gavage at doses of 37, 75 or 150 mg kg^{-1} body weight, 5 days per week for 105 weeks. A vehicle control group of 50 male rats received 0.5% methylcellulose only. The 2-year survival rates in the control group and the 37, 75 and 150 mg kg^{-1} dose groups were 40%, 33%, 32% and 0% respectively. A survival-adjusted analysis is indicated by the dose-related increases in mortality.

We were interested in skin fibromas, which are occult tumours of the subcutaneous tissue. The data are given in Table 1. As mentioned earlier, the observed incidences in the four groups were 1, 9, 8 and 5, and the poly-3 survival-adjusted rates were 2.4%, 22.3%, 20.6% and 15.3% respectively. Thus, even after adjusting for differences in survival, there was an apparent downturn in skin fibroma rates with increasing dose levels of methyleugenol.

The NTP concluded that methyleugenol showed clear evidence of carcinogenic activity in male rats. This decision took into consideration the increases in several types of tumour, including skin fibromas. Typically, the poly-3 adjusted CA trend test plays a key role in the decision-making process. In the case of skin fibromas, however, the NTP had to rely on pairwise comparisons and informal scientific judgment because the trend test was not statistically significant, i.e. using the CA_{P_3} test to compare separately each dose group with the control group gave p-values of 0.006, 0.011 and 0.055, but the overall CA_{P_3} test for trend was not significant ($p = 0.108$), presumably because of the downturn in rates at the top dose. Alternatively, the proposed test W_{P3} yields evidence of a statistically significant upward trend in the incidence of skin fibromas with increasing doses of methyleugenol ($p = 0.027$). Since the historical control rate of this neoplasm in comparable studies was reported by the NTP (National Toxicology Program, 2000) to be only 4.2% (17/402), with a range of 0–12%, incidences as high as those observed in the dosed groups are unlikely to be due to chance. Thus, our test formally produces results which are consistent with the pairwise comparisons and the NTP's scientific judgment, despite the differential mortality and the downturn at the top dose.

5. Simulation study

The goal of our simulation study is to demonstrate that the new test procedure performs better than the popular poly-3 adjusted CA test in terms of size and power. To accomplish this, we simulated representative data from a broad collection of situations that are commonly encountered in rodent bioassays. In the following subsections, we describe the design and results of our simulation study.

5.1. Study design

As this paper focuses on methods for analysing rodent carcinogenicity studies, we present simulations which are modelled after experiments that were conducted by the US NTP. The duration of a typical NTP rodent bioassay is 2 years; thus, we simulate data from a study that ends with a sacrifice at 24 months ($t_s = 24$). For each animal, we generate two random variables: T_1, which is the time to onset of the tumour, and T_2, which is the time to death from natural causes. An irreversible tumour is present at death if and only if $T_1 < \min(T_2, t_s)$, where the minimum of T_2 and t_s is the observed time of death. The animal is killed at the terminal sacrifice if $t_s < T_2$ and otherwise dies of natural causes.

We simulated data from $k = 4$ dose groups. Let $\mathbf{d}' = (d_1, d_2, d_3, d_4)$, where the control dose is $d_1 = 0$. For the ith dose group, we assume that T_{i1} and T_{i2} are independent Weibull random variables with distributions

$$P(T_{ij} > t | d_i) = \exp\{-(\psi_j \phi_{ij}^{d_i}) t^{\gamma_j}\}$$

for $j = 1, 2$. The shape parameters γ_1 and γ_2 govern the steepness of the tumour incidence and mortality curves respectively. We assume that the shape parameters do not vary with the dose, but that the scale parameters $\psi_j \phi_{ij}^{d_i}$ are related to the dose through a log-linear model. Thus, ψ_1 and ψ_2 are base-line scale parameters whose inverses represent the average time to tumour onset and time to death from natural causes respectively in the control group ($i = 1$). Finally, we assume a common scale multiplier for mortality, $\phi_{i2} \equiv \phi_2$, and focus on the ratio of the tumour incidence rate in the ith dose group relative to the control group, say $\theta_i = \phi_{i1}^{d_i}$.

Historical information from past NTP studies was used to select reasonable values for the simulation parameters. As our focus is on incidence, we evaluated more tumour onset patterns than mortality patterns. In fact, we chose a single (typical) mortality curve for the control group, though we considered several dose effects on that base-line mortality. We fixed the mortality shape parameter at $\gamma_2 = 5$, and we chose a base-line scale parameter of $\psi_2 = 4.479 \times 10^{-8}$ so that control survival at 2 years would be 70%, a rate which is often observed in practice. Dose affects base-line mortality through the scale multiplier $\phi_2^{d_i}$. We considered two dose patterns that are commonly seen in NTP studies: $\mathbf{d}' = (0, 0.5, 1, 2)$ and $\mathbf{d}' = (0, 0.1, 0.5, 2.5)$, which we refer to as having twofold and fivefold spacings respectively. We investigated five choices for ϕ_2: 1, 1.5, 2, 2.5 and 3, which produced dose effects on mortality ranging from 'none' ($\phi_2 = 1$) to 'severe' ($\phi_2 = 3$). Similarly to the methyleugenol study, the severe dose effect corresponds to a 2-year survival rate in the highest dose group that is close to 0%.

Regarding incidence, we evaluated three values for the tumour onset shape parameter: $\gamma_1 = 1.5, 3, 6$. The middle value ($\gamma_1 = 3$) is optimal for the CA_{P_3} test, whereas the other two values were included to study the robustness of the CA_{P_3} and W_{P3} tests to violations of the poly-3 shape assumption. Next, we considered four values for the proportion of tumour-bearing animals in the control group: $\pi_1 = 0.01, 0.05, 0.20, 0.30$. The smallest value reflects a 'rare' tumour and the largest value reflects a 'common' tumour. Then, for fixed π_1, γ_1, γ_2 and ψ_2, we used equation (1) to solve for ψ_1. The respective values for ψ_1 associated with the four values of π_1 are 0.009×10^{-2}, 0.047×10^{-2}, 0.206×10^{-2} and 0.330×10^{-2} when $\gamma_1 = 1.5$, 0.008×10^{-4}, 0.042×10^{-4}, 0.185×10^{-4} and 0.297×10^{-4} when $\gamma_1 = 3$, and 0.007×10^{-8}, 0.033×10^{-8}, 0.144×10^{-8} and 0.232×10^{-8} when $\gamma_1 = 6$. Finally, for each of the two dose patterns, we chose ϕ_{i1}-values that gave the following six patterns for the incidence ratio, $\theta' = (\theta_1, \theta_2, \theta_3, \theta_4)$: $(1, 1, 1, 1)$, $(1, 1, 1, 4)$, $(1, 1, 4, 4)$, $(1, 4, 4, 4)$, $(1, 2, 2, 4)$ and $(1, 2, 3, 4)$. The null hypothesis corresponds to $\theta' = (1, 1, 1, 1)$.

Thus, we investigated 720 sets of simulation parameters by taking all combinations of two dose patterns, five mortality scale multipliers, three incidence shapes, four control tumour rates

and six incidence ratio patterns. Type I error rates were evaluated for the 120 configurations associated with the null hypothesis, and powers were estimated for the 600 configurations associated with various alternative hypotheses. For each configuration, we compared the proposed trend test (W_{P3}) with the poly-3 adjusted CA trend test (CA_{P_3}). Every simulation experiment was based on 10000 runs and a nominal level of 5%.

5.2. Results

The key result is that the W_{P3} test generally makes fewer (and smaller) type I errors than does the CA_{P_3} test, and yet W_{P3} is more powerful in the majority of the situations that were investigated. As one way of summarizing the operating characteristics of the two tests, Fig. 1 plots the size estimates for W_{P3} *versus* CA_{P_3} in all 120 null cases (Figs 1(a) and 1(b)) and the power estimates in all 600 non-null cases (Figs 1(c) and 1(d)). Figs 1(a) and 1(c) correspond to the twofold dose spacing, and Figs 1(b) and 1(d) correspond to the fivefold dose spacing. Diagonal reference lines are drawn to facilitate comparisons of the tests. Furthermore, the size plots are subdivided into quadrants by horizontal and vertical lines drawn at 0.055, which is the cut point that we use for classifying a test as liberal (i.e. anticonservative) or not. This cut point corresponds to the nominal level (5%) plus two standard errors, rounded up to the nearest tenth of a per cent.

First, consider the size results that are depicted in Figs 1(a) and 1(b). The bulk of the symbols lie near but below the diagonal lines. Thus, the two tests operate at about the same level in most cases, though the rejection rates tend to be a little lower for W_{P3} than for CA_{P_3}. The lower left-hand quadrants illustrate that both tests operate at (or below) the correct level in the majority (79/120) of the configurations that were considered. Conversely, the upper right-hand quadrants show that both tests are liberal in some situations, though W_{P3} is always less liberal than CA_{P_3}. A closer examination reveals that the incidence shape parameter γ_1 was small (i.e. the open squares) in all 22 null configurations for which both tests were liberal. The upper left-hand quadrants are completely empty, which means that W_{P3} was never liberal when CA_{P_3} was not. In contrast, however, the lower right-hand quadrants show that there are 19 configurations for which CA_{P_3} was liberal whereas W_{P3} maintained the correct level. This shortcoming of CA_{P_3} relative to W_{P3} was highly correlated with the dose spacings being wide (i.e. fivefold).

In addition to the magnitude of the incidence shape parameter, the phenomenon of both tests being liberal is strongly correlated with differential mortality across dose groups. The smaller the shape parameter and the greater the dose effect on mortality, the more likely both tests are to be liberal. Kodell *et al.* (1994) made a similar observation regarding the size of the CA_{P_3} test; in a simulation with $\gamma_1 = 1$, they found that CA_{P_3} rejects increasingly too often as the dose effect on mortality increases. Conventional wisdom says that almost any test performs well when survival is the same across dose groups, but our simulation shows that this is not necessarily true for rare tumours and wide dose spacings (on the log-scale). For example, in three of our null configurations for which tumours are rare (i.e. $\pi_1 = 1\%$), the type I error rate for CA_{P_3} exceeds 9.5%, even though there is absolutely no dose effect on mortality. In fact, the size of the CA_{P_3} test is 10.3% in one of these cases, even though the shape parameter γ_1 is exactly 3. Kodell *et al.* (1994) did not notice this effect because they examined only background tumour rates as small as 5%. In contrast, the size of the proposed W_{P3} test never exceeded the nominal level in the absence of differential mortality, regardless of the incidence shape parameter or the background tumour rate. Similarly, W_{P3} was never liberal for $\gamma_1 = 3$ and $\gamma_1 = 6$, regardless of the dose effect on survival, the spacing of the doses or the background tumour rate.

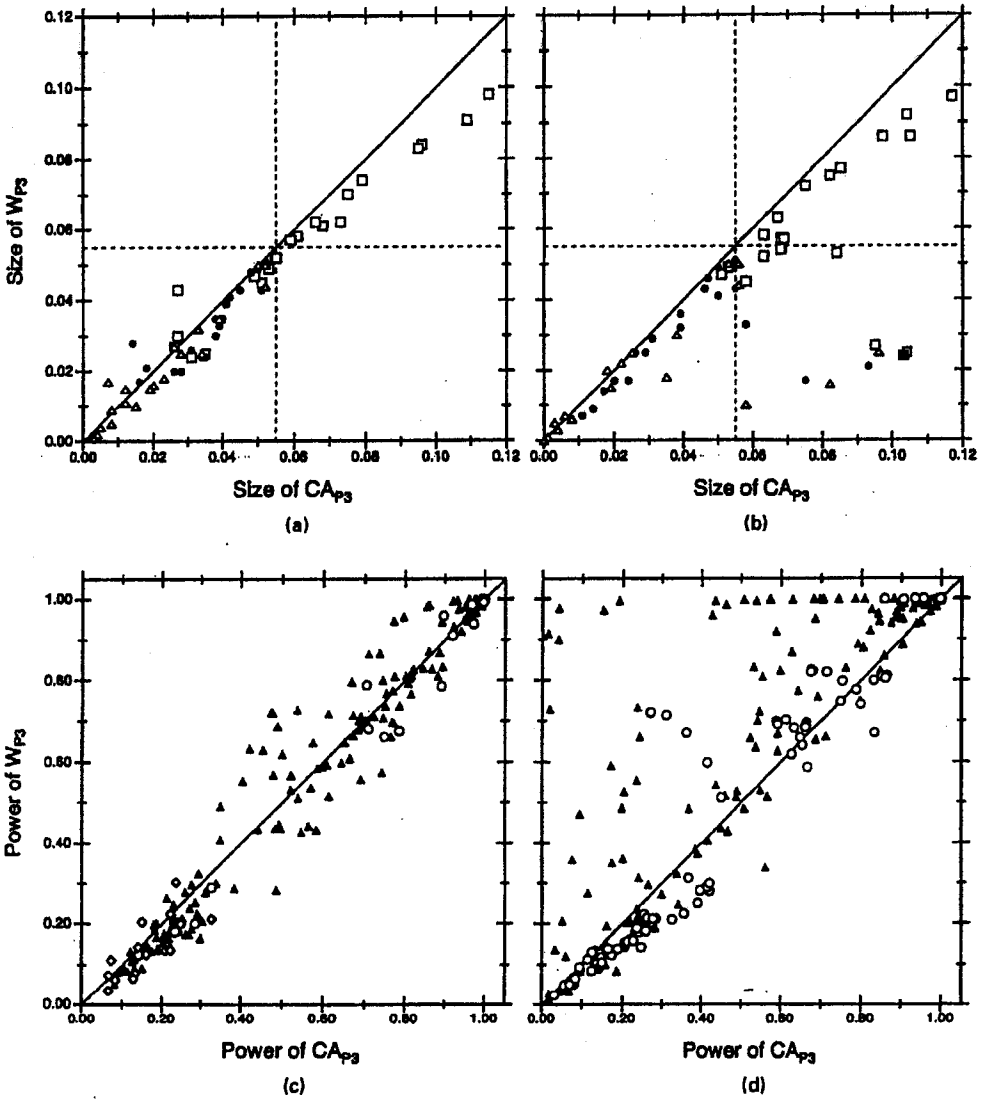

Fig. 1. (a), (b) Size and (c), (d) power comparisons for CA_{P_3} and W_{P3} based on 10000 simulated data sets (□, •, △, estimated sizes when the incidence shape parameter γ_1 is 1.5, 3 and 6 respectively; ○,▲,○, estimated powers when the size of W_{P3} is less than, equal to (within 0.01) and greater than the size of CA_{P3} respectively; ¦, - - - - - - -, markers used to classify the tests as liberal or not); (a), (c) simulation configurations with twofold dose spacing; (b), (d) simulation configurations with fivefold dose spacing

The test proposed also outperformed the conventional test with respect to power. Overall, W_{P3} was at least as powerful as CA_{P_3} in 59% of the 600 non-null configurations, which is remarkable considering that CA_{P_3} had the larger size in 83% of the null cases. To allow a more meaningful assessment, we adjusted for each test's type I error rate, as the test with the larger size otherwise might be expected to have a power advantage. We formed three subsets, according to whether the size of W_{P3} was less than, equal to (i.e. within 0.01) or greater than the size of

CA_{P_3}. The numbers of configurations in these subsets were 155, 430 and 15 respectively. Figs 1(c) and 1(d) plot the estimated powers, using distinct symbols for each of the size-specific subsets. Among the 430 cases for which the two tests operated at the same level, W_{P3} was more powerful than CA_{P_3} in 65% of these situations. In fact, W_{P3} was over 60 times more powerful than CA_{P_3} in one case (91.4% *versus* 1.5%), whereas the power of CA_{P_3} was never much larger than twice that of W_{P3} (18.6% *versus* 8.4%), even among the 155 configurations for which CA_{P_3} had a higher type I error rate. Clearly, the power advantages of W_{P3} were generally greater in number and magnitude than those of CA_{P_3}.

The dose spacing was perhaps the most important factor in determining when the proposed test was more powerful than the conventional test. Motivated by dose patterns that have been seen in typical rodent tumourigenicity studies, such as the NTP bioassays, our simulation considered a doubling and a quintupling of dose levels. For twofold dose spacings, both tests had approximately the same power, as seen by the proximity of most of the symbols in Fig. 1(c) to the diagonal line. For fivefold dose spacings, however, W_{P3} often had much higher power than CA_{P_3}; virtually all the symbols in Fig. 1(d) that were not reasonably close to the diagonal were in the upper triangular region. The largest power advantages of W_{P3} correspond to configurations where both tests have equal sizes, but there are some situations in which W_{P3} was more powerful even when its size was smaller (i.e. the open circles). The key result, as illustrated by Figs 1(c) and 1(d), is that W_{P3} is nearly always as powerful as CA_{P_3}, and in some situations W_{P3} is much more powerful than CA_{P_3}.

6. Conclusions

In the process of analysing the skin fibroma data from the NTP methyleugenol study, we developed a simple, yet powerful, alternative to the commonly used poly-3 adjusted CA (CA_{P_3}) trend test. Our new procedure is oriented towards monotone tumour incidence rates, but it does not assume a particular model (or family of models) for the dose–response curve; nor does it depend on the numerical scores that are assigned to the dose groups. In contrast, not only does the CA_{P_3} test depend on the dose scores, but it is also oriented towards tumour proportions that are linear in dose. This linearity assumption is not valid in general, especially when the treatment has a strong effect on survival, such as in the methyleugenol study. Dose-related mortality can produce a downturn in the proportions of tumour-bearing animals at the highest dose(s), even when the incidence rates are monotone in dose. Thus, the power of the CA_{P_3} test can be extremely low in these situations, whereas the power of the test proposed should not suffer.

The NTP currently uses a form of the CA_{P3} test to assess tumorigenicity in their 2-year cancer bioassays. We showed that the proposed test W_{P3} consistently outperformed the conventional CA_{P3} test with respect to both size and power. Thus, our results should be relevant and important to the NTP, in particular, and interesting in general to a broad audience of researchers who are involved in the analysis of carcinogenicity data.

Acknowledgements

We appreciate the comments of David Dunson, Richard Morris, the Joint Editor and a referee which have improved the presentation of the paper.

References

Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.
Bailer, A. and Portier, C. (1988) Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics*, **44**, 417–431.

Barlow, R., Bartholomew, D., Bremner, J. and Brunk, H. (1972) *Statistical Inference under Order Restrictions*. London: Wiley.

Bieler, G. and Williams, R. (1993) Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. *Biometrics*, **49**, 793–801.

Cochran, W. (1954) Some methods for strengthening the common χ^2 tests. *Biometrics*, **10**, 417–451.

Dinse, G. (1991) Constant risk differences in the analysis of animal tumorigenicity data. *Biometrics*, **47**, 681–700.

Dinse, G. (1998) Tumour incidence experiments. In *Encyclopedia of Biostatistics*, vol. 6 (eds P. Armitage and T. Colton), pp. 4597–4609. Chichester: Wiley.

Dunson, D. B. and Dinse, G. E. (2001) Bayesian incidence analysis of animal tumorigenicity data. *Appl. Statist.*, **50**, 125–141.

Kodell, R., Blackwell, B., Bucci, T. and Greenman, D. (1995) Cause-of-death assignment at the National Center for Toxicological Research. *Toxicol. Pathol.*, **23**, 241–247.

Kodell, R., Chen, J. and Moore, G. (1994) Comparing distributions of time to onset of disease in animal tumorigenicity experiments. *Commun Statist. Theory Meth.*, **23**, 959–980.

Mancuso, J., Ahn, H. and Chen, J. (2001) Order-restricted dose-related trend tests. *Statist. Med.*, **20**, 2305–2318.

Mancuso, J., Ahn, H., Chen, J. and Mancuso, J. (2002) Age-adjusted exact trend tests in the event of rare occurrences. *Biometrics*, **58**, 403–412.

McKnight, B. and Crowley, J. (1984) Test for differences in tumor incidence based on animal carcinogenesis experiments. *J. Am. Statist. Ass.*, **79**, 639–648.

National Toxicology Program (2000) Toxicology and carcinogenesis studies of methyleugenol in F344/N rats and B6C3F$_1$ mice. *Technical Report 491*. US Department of Health and Human Services, National Institutes of Health, Washington DC.

Peddada, S., Prescott, K. and Conaway, M. (2001) Tests for order restrictions in binary data. *Biometrics*, **57**, 1219–1227.

Peto, R., Pike, M., Day, N., Gray, R., Lee, P., Parish, S., Peto, J., Richards, S. and Wahrendorf, J. (1980) Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments. In *Long-term and Short-term Screening Assays for Carcinogens: a Critical Appraisal*, annex to suppl. 2, pp. 311–426. Lyon: International Agency for Research on Cancer.

Williams, D. (1977) Some inference procedures for monotonically ordered normal means. *Biometrika*, **64**, 9–14.